BILINGUAL PUBLISHING CO.
Pioneer of Global Academics Since 1984

Journal of Computer Science Research
https://ojs.bilpublishing.com/index.php/jcsr

ARTICLE

A Mathematical Theory of Big Data

Zhaohao Sun*

Department of Business Studies, Papua New Guinea University of Technology, Lae 411, Morobe, Papua New Guinea

ARTICLE INFO

Article history
Received: 21 April 2022
Accepted: 19 May 2022
Published Online: 31 May 2022

Keywords:
Big data
Big data analytics
Fuzzy logic
Similarity
Discrete mathematics

ABSTRACT

This article presents a cardinality approach to big data, a fuzzy logic-based approach to big data, a similarity-based approach to big data, and a logical approach to the marketing strategy of social networking services. All these together constitute a mathematical theory of big data. This article also examines databases with infinite attributes. The research results reveal that relativity and infinity are two characteristics of big data. The relativity of big data is based on the theory of fuzzy sets. The relativity of big data leads to the continuum from small data to big data, big data-driven small data analytics to become statistical significance. The infinity of big data is based on the calculus and cardinality theory. The infinity of big data leads to the infinite similarity of big data. The proposed theory in this article might facilitate the mathematical research and development of big data, big data analytics, big data computing, and data science with applications in intelligent business analytics and business intelligence.

1. Introduction

Big data has become one of the most important frontiers for innovation, research, and development in data science, computer science, artificial intelligence, industry, and business [1,2]. Big data has also become a strategic asset for nations, organizations, industries, enterprises, businesses, and individuals [3,4]. Big data technology including big data analytics has been successfully used to explore business insights and data intelligence from big data [2,5]. Mathematics researchers have paid increasing attention to the dramatic development of big data and its impact on mathematics by offering courses and holding workshops to develop the mathematics of big data [6,7]. However, there is no literature on a mathematical theory of big data based on the search using Google Scholar and Scopus (accessed on April 28, 2022). This indicates that a mathematical theory of big data has lagged far behind the big intelligence, big service, and big market opportunity resulting from big data [8]. The above brief analysis implies that the followings are still big issues for big data toward the establishment of a mathematical theory.

- What is a mathematical theory of big data?
- How does a social networking platform become an outstanding contributor to big data?

This article addresses these two issues by providing a mathematical theory of big data. This mathematical theo-

*Corresponding Author:
Zhaohao Sun,
Department of Business Studies, Papua New Guinea University of Technology, Lae 411, Morobe, Papua New Guinea;
Email: zhaohao.sun@pnguot.ac.pg; zhaohao.sun@gmail.com

DOI: https://doi.org/10.30564/jcsr.v4i2.4646

ry covers the cardinality approach, fuzzy logic approach, similarity approach, and logical approach due to the space limitation. More specifically, it proposes "big" as an operation and presents a cardinality approach to the big volume of big data, the latter is one of the ten big characteristics of big data [5], a fuzzy logic-based approach to big data, a similarity-based approach to big data, and a logical approach to the marketing strategy of big data-driven social networking services.

The remainder of this article is organized as follows. Section 2 presents a cardinality approach to big data. Section 3 looks at searching for big data using the set theory. Section 4 applies a fuzzy-logic approach to big data from a relativity perspective. Section 5 proposes a similarity-based approach to big data. Section 6 looks at a logical approach for promoting big data-driven social networking services. The final sections discuss the related work and end this article with some concluding remarks and future work.

It should be noted that this article does not directly address the issues related to how to efficiently manage, analyze, mine, and process big data although the proposed mathematical theory can be applied to each of them.

2. A Cardinality Approach to Big Data

This section examines cardinality of big data as a mathematical approach to the big volume of data based on discrete mathematics and cardinal number theory in real analysis. It addresses what the biggest volume of big data is.

Definition 1

Let S be a set. The cardinality of S is m, denoted by $|S| = m$, if there are exactly m distinct elements in S, where m is a non-negative integer [9].

Example 1.

Let $S = \{a, b, c, d, e, f, g\}$, then $|S| = 7$.

S is said to be finite if m is a non-negative integer. Otherwise, S is said to be infinite. The cardinality of an infinite set S is discussed in set theory and discrete mathematics [10-12].

A large number of research articles on big data have been published in the past decade since 2012 [8,13-16]. Most of them consider volume as the first V of big data [8]. They mentioned terabytes (TB, 2^{40} B), petabytes (PB, 2^{50} B), exabytes (EB, 2^{60} B)[16], zettabytes (ZB, 2^{70} B), and yottabyte (YB, 2^{80} B) as the big volume of big data. Google, YouTube, and other global data giants have the volume of big data at a PB level yearly, while Amazon has big data at an EB level yearly [17]. It seems that the principle of big data is that the bigger volume the big data has, the more important it is. This is true in some cases, for example, Google and Amazon with a bigger volume of big data have a big value in the market.

However, when the author asked his friend's child to count numbers, 1, 2, 3, ..., then only a few minutes later, he could not like to count numbers anymore. Then he said "infinity" as the conclusion of his counting number. What an interesting child he is! He intuitively knew the end of counting numbers is infinity. The generalization of this story is the cardinality of big data.

In entity-relationship modeling, an attribute value is the least unit for representing data [18]. An attribute value has also been the least unit for defining and manipulating data in database systems [19]. An attribute value is still the least unit for NoSQL database or web data processing. Therefore, the attribute value can be used as the least unit of big data. An attribute value can be denoted as v. For example, $v_1 = big$, $v_2 = data$, $v_3 = analytics$, $v_4 = intelligence$, $v_5 = service$, etc. These values can be considered as keywords when searching online and or stop words in natural language processing. From a linguistic viewpoint, they are the elements for constructing a sentence, a paragraph, a text, and so on. However, for searching, some attribute values, for example, *a*, *an*, *the*, *of*, can be considered negligible components. An attribute value can be any word(s) (e.g., in English) in the web text. Using number sequence and limit in mathematics [20], an attribute value sequence is $v_1, v_2, \cdots, v_n, \cdots$, and n is an integer. Now a question arises,

What does the limit of attribute value sequence v_i mean when i tend to infinity?

Let V be a set of attribute values, and U be the universe of big data. U consists of all online and offline data available to mankind. Therefore, U includes all the data, information, knowledge, experience, intelligence, and wisdom in either article or website, or multimedia form [9]. Then, the relationship between V and U is as follows.

$$V \subseteq U \tag{1}$$

In other words, V is a subset of U.

A finite attribute value sequence $v_1, v_2, \cdots, v_n$ can be used to constitute a sentence using concatenation, where n is an integer. However, from a perspective of human cognition [21], n is not a fixed integer. The corresponding attribute value sequence $v_1, v_2, \cdots, v_n$ cannot represent all the data and knowledge existing in the world. At least a countable infinite number of attribute values is required to constitute all the big data, big information, big knowledge, big intelligence, and big wisdom [22], because the number of English sentences is theoretically infinite [23]. This implies that the cardinality of V is $|V| = \aleph_0$. $\aleph_0$ is the cardinality of all integers N, then U should be at least uncountable infinity as the cardinality of all the real numbers [24], because any subset of V can be constituted to a meaningful sentence, paragraph, or text. It can also correspond to

at least a picture or a set of pictures, such as a data stream. For example, if one searches "Paul" (as a v_1), then one will find a set of texts or pictures consisting of "Paul", even using Google (see the next section). This means that an element of V corresponds to a set of elements of U. Therefore, a relationship between the cardinality of V and that of U is

$$|U| = 2^{|V|} \tag{2}$$

and

$$|U| = 2^{|V|} = 2^{\aleph_0} = c \tag{3}$$

where c is the cardinality of the real number set. Strategically, if one likes to understand the existing finite world of big data, one should "live" in the infinite world of big data. It is important to be a follower in the finite world of big data in terms of EB, ZB, and YB. It is also important to be an explorer in the infinite world of big data. For the former, we enjoy the 3G communication using a mobile phone in the 2000s, whereas for the latter one should look at what will be the next generation of smartphones using 6G or 7G communication.

It should be noted that this section is motivated by a large number of articles or books using petabytes, exabytes, and zettabytes, as well as yottabytes, to describe how big the volume of big data [17,25,26]. Therefore, it is interesting to answer how big the volume of big data is in the future. The cardinality theory [12,24] is used to develop it in some detail. In other words, this section provides an answer to the question: how big is the volume of big data eventually, using real analysis or measure theory.

3. A Set Theory for Searching Big Data

This subsection discusses searching for big data using the set theory, which is the foundation of modern mathematics [12,24].

Let $u \in U$ be a document on the web. u may be a Microsoft word file in .docx or report in pdf. Let $v \in V$ be an attribute value. v may be a word such as "data", or "intelligence", or "wisdom", or "engineering", then a search function denoted as $s: V \to U$, is defined as

$$s(v) = u, \text{ if } v \in u \tag{4}$$

For example, if one uses Google Scholar to search for "big data", denoted as v, then she or he finds 1,750,000 results (retrieved on April 26, 2022), denoted as u, each of them should include v.

More generally, a search function can be defined as

$$s(v) = F(v) \tag{5}$$

where $F(v) = \{u_i | \ v \in u_i, u_i \in U, i = \{0,1,2,\cdots,m\}\}$; $i = 0$ means that "no research results" for v. This is valid for searching practice using all the search engines online including Google, Baidu, Semantic Scholar, and Google Scholar. The core idea behind the online search is that one keyword search corresponds to at least a picture/document or a set of pictures/documents as the search results. A Google search for "big data" found 61,700,000 results (retrieved on 22 April 2018) and 398,000,000 results (retrieved on 20 April 2022). Therefore, $F(v) \subseteq U$.

Searching on the web using search engines such as Google and Baidu is an operation. Search or query in a relational database using SQL (Structured Query Language) is a data operation (data manipulation) [19]. SQL should be renamed as Structured Query Engine (SQE) based on the usage of the search engine. At least the author discussed it with his colleagues. In what follows, this section looks at the property of the search function as an operation.

Let $v_1, v_2, v_3 \in V$, using Equation (5), $s(v_1) = F(v_1)$, $s(v_2) = F(v_2)$, $s(v_3) = F(v_3)$. Then the following property of search functions holds [10].

$$s(v_1 \vee v_2) = s(v_1) \cap s(v_2) = F(v_1) \cap F(v_2) \tag{6}$$

where $\vee$ is a space operation between v_1 and v_2 to reflect the search using Google and Baidu. $\vee$ as an operator has the property of association, that is, $v_1 \vee (v_2 \vee v_3) = (v_1 \vee v_2) \vee v_3$ [27]. $\vee$ is similar to concatenation between data items in linguistics or formal language [11].

Now given an attribute value sequence $v_1, v_2, \cdots, v_n, \cdots$, then $s(v_1) = F(v_1)$, $s(v_2) = F(v_2), \ldots$, $s(v_n) = F(v_n)$, ...

Theorem 1.

In the finite world of big data, the search result with respect to operation $\vee$ is

$$s(v_1 \vee v_2 \vee \cdots \vee v_n) = \prod_1^n F(v_i) = F(v_1) \cap F(v_2) \cap \cdots \cap F(v_n) \tag{7}$$

Theorem 2.

When $n \to \infty$, the following property of search as an operation $\vee$ in the web search holds.

$$s(v_1 \vee v_2 \vee \cdots \vee v_n \vee \cdots) = \prod_1^\infty F(v_i) \tag{8}$$

Theorem 1 and Theorem 2 can be proved based on Equation (6) easily.

Equation (7) and Equation (8) are representation theorems for searching in the finite world and infinite world of big data respectively.

Many people have the experience of searching for what they expect on the web using Equations (6 and 7), although each of them has not experienced the search of the web based on the Equation (8). This is because an individual's search on the web is finite (in terms of attribute value) whereas all the human being's searches on the web should be infinite.

Based on the dual principle of the set theory, the $\vee$ operation of $v_1 \vee v_2$ motivates us to introduce $\wedge$ operation [27].

Let's look at the following example: Paul just searched for "intelligent big data analytics" using Google Scholar. However, he had not searched for what he expected, so he had to extend his search space by using "big data analytics". Let "intelligent big data analytics" be v_1 and "big data analytics" be v_2. Then Paul's extending search space means that he uses $v_1 \wedge v_2$ (a kind of intersection in set theory [24]) to search for what he expected on the web. Then the following two theorems hold corresponding to Equations (7) and (8), based on the dual principle of set theory [9].

Theorem 3.

In the finite world of big data, the search results with respect to operation $\wedge$ are

$$s(v_1 \wedge v_2 \wedge \cdots \wedge v_n) = \coprod_1^n F(v_i) = F(v_1) \cup F(v_2) \cup \cdots \cup F(v_n) \quad (9)$$

Theorem 4.

When $n \to \infty$, the following property of search as an operation $\wedge$ in the web search holds.

$$s(v_1 \wedge v_2 \wedge \cdots \wedge v_n \wedge \cdots) = \coprod_1^{\infty} F(v_i) \quad (10)$$

Equations (7), (8), (9), and (10) are a mathematical basis for searching for big data on the web.

4. A Fuzzy Logic Approach to Big Data

Fuzzy sets theory has been successfully applied in many areas including finance, database, pattern recognition, and natural language processing, to name a few [28,29]. Fuzzy logic can be applied to address the vagueness and veracity of big data [14,29,30]. This section uses fuzzy sets and fuzzy logic to examine the relativity of big data as a fundamental of big data.

Let U be the universe of big data, and $n \in N$. Then big as an attribute value is an element of U, that is, {big} $\in U$.

A fuzzy set of big in N is defined with a membership (characteristic) function $f_{big}(n)$ which associates every number of $n \in N$ with a real number in the interval [0,1] [30], that is,

$$f_{big}(n) \in [0,1]$$

If $f_{big}(n) = 1$, $n \in N$ is said to be big, otherwise, $f_{big}(x) < 1$, $n \in N$ is said to be not big. In fuzzy logic, "not big" does not mean "small" [28].

A question arises: What big does mean in big data from a perspective of fuzzy logic? To answer this question, this subsection examines an average child at 5 years old, a young person at 20 years old, and a graduate with a Bachelor of Data Science, and observes what they believe big as a term [9].

For the child, he believes that 5,000 is big, denoted as $f_{big}(n, p_1) \in [0,1]$, because he likes to have US$ 5,000, then

$$f_{big}(n, p_1) = \begin{cases} 1, & if\ n \geq 5{,}000 \\ < 1, & otherwise \end{cases}$$

This equation indicates that the child believes that any number greater than 5,000 will be "big", whereas number less than 5,000 is not big.

For the young person, he believes that 1 million is big, denoted as $f_{big}(n, p_2) \in [0,1]$, because he likes to be a millionaire, that is,

$$f_{big}(n, p_2) = \begin{cases} 1, & if\ n \geq 1{,}000{,}000 \\ < 1, & otherwise \end{cases}$$

This means that the young person believes that any number greater than 1,000,000 will be "big", whereas number less than 1,000,000 is not big.

For the graduate with the degree, he believes that 1 billion is big, denoted as $f_{big}(n, p_3) \in [0,1]$, because he likes to be a billionaire, that is,

$$f_{big}(n, p_3) = \begin{cases} 1, & if\ n \geq 10^9 \\ < 1, & otherwise \end{cases}$$

In other words, he believes that any number greater than 10^9 will be "big", whereas number less than 10^9 is not big.

The above analysis using fuzzy sets indicates that all persons unanimously agree that a number less than 5,000 is "not big", and different people have a different understanding of big as the term. However, all people unanimously have a concept of "big" in numbers motivated by their backgrounds, environments, and expectations.

Let $P = \{p_k | p_k\ is\ a\ person, k = 1, 2 \cdots, m\}$. $m \in N$ is a given natural number, it can be the total number of all the people living in the world. For every person, p_k, his or her perspective to big can be represented as a fuzzy set Bp_k with the following membership function $f_{big}(n, p_k) \in [0,1]$ and

$$f_{big}(n, p_k) = \begin{cases} 1, if\ n \geq n_k \\ < 1, otherwise \end{cases} \quad (11)$$

For example, the above child can be named a p_1, then the perspective of p_1 to "big" satisfies the following properties: $n_1 = 5{,}000$.

$$f_{big}(n, p_1) = \begin{cases} 1, if\ n \geq n_1 \\ < 1, otherwise \end{cases}$$

Based on the operations of fuzzy sets [28,30], the intersection of the all fuzzy sets Bp_k ,$k = 1, 2 \cdots, m$ with membership functions like Equation (12) is a fuzzy set C, written as $C = Bp_1 \cap Bp_2 \cap \cdots \cap Bp_M$, whose membership function $f_{big}(n, C)$ is

$$f_{big}(x, C) = Min(f_{big}(n, p_1), f_{big}(n, p_2), \cdots, f_{big}(n, p_m)) \quad (12)$$

Or, in abbreviated form

$$f_{big}(n, C) = \wedge_1^m (f_{big}(n, p_k)) \quad (13)$$

The membership function $f_{big}(n, C)$ indicates that there exists a number $K \in N$ such that for every $p_k, k = 1, 2 \cdots, m$.

$$f_{big}(n, p_k) = \begin{cases} 1, & when\ n \geq K \\ 0, & otherwise \end{cases} \quad (14)$$

where $K = \max(n_1, n_2, \cdots, n_m)$. In other words, all the people have unanimously agreed that *K* or greater is big. In terms of big data, this means that all the people have unanimously agreed that the data with *K*Bytes or greater is big data. This result conforms to the currently popular idea in the literature of big data, that is, data with Exabytes or Zettabytes are big data [13,31], whereas data with an MB cannot be considered big data anymore, although it used to be big data two decades ago [9].

The union of the all-fuzzy sets $Bp_k, k = 1, 2 \cdots, m$ with membership functions in Equation (11) is a fuzzy set *D*, written as $D = Bp_1 \cup Bp_2 \cup \cdots \cup Bp_m$, whose membership function $f_{big}(x, D)$ is

$$f_{big}(n, D) = Max(f_{big}(n, p_1), f_{big}(n, p_2), \cdots, f_{big}(n, p_m)) \quad (15)$$

Or, in abbreviated form [30]

$$f_{big}(n, D) = \vee_1^m (f_{big}(n, p_k)) \quad (16)$$

The membership function $f_{big}(n, D)$ indicates that there exists a number $H \in N$ such that for every $p_k, k = 1, 2 \cdots, m$.

$$f_{big}(n, p_k) = \begin{cases} 1, & when\ n \geq H \\ 0, & otherwise \end{cases} \quad (17)$$

where $H = \min(n_1, n_2, \cdots, n_m)$. Different from the above analysis based on the intersections of fuzzy sets, the union of fuzzy sets [28] indicates that for any given $J \in N$, if there exists a person who believes that J is big, that is, $J \geq H$, then all the other persons have to agree that J or greater is big. In terms of big data, this means that if there exists a person who believes that data with $Jbytes$ is big data, then all the other persons have to agree that the data with $Jbytes$ or greater is big data. If this is true, then J might be $100 \in N$, because a child might consider 100 as a big number. In other words, the statement that the data with Exabytes or Zettabytes is just big lacks evidence from the social reality based on the theory of fuzzy logic.

The above discussion implies that one characteristic of big data is relativity, that is, big is a relativity concept; the relativity of big data is a fundamental of big data. The secret behind the relativity of big data is inclusiveness, that is, we have to permit that every individual has his or her understanding of what big means in big data [9]. This inclusiveness can make big more powerful in terms of research and development of big data. This relativity of big data also brings forth that a universal benchmark does not exist for big volume, big variety, big velocity, and big veracity that define and measure the characteristics of big data [14].

Finally, the relativity of big data leads to the continuum from small data to big data, big data-driven small data analytics is of statistical significance [14].

5. A Similarity-based Approach to Big Data

The concept of similarity is a fundamental concept in mathematics, computer science, data science, and artificial intelligence [2,10]. For example, "similar problems have similar solutions" is the principle of case-based reasoning [32]. In other words, case-based reasoning is a process of discovering similarity intelligence from a case base, just as data mining is a process of discovering data intelligence from a database or big data [22,26].

Definition 2.

A binary relation *S* on a non-empty set *X* is a similarity relation if it satisfies [32]

(R) $\forall x, xSx$;

(S) If xSy, then ySx;

(T) If xSy and ySz, then xSz.

The conditions (R), (S), and (T) are the reflexive, symmetric, and transitive laws. If xSy or $< x, y > \in S$ then x and y are said to be similar, denoted as $x \approx y$.

Based on this definition, this section introduces finite similarity, infinite similarity, weak similarity, strong similarity, and limit of similarity. All these concepts of similarity are important for similarity-based reasoning for big data to discover similarity intelligence and data intelligence from big data.

5.1 Finite Similarity and Infinite Similarity

From a viewpoint of search online, a search engine cannot limit the number of text words. Different texts have different lengths in words. Therefore, the search space is infinite and consists of infinite texts or words.

As mentioned earlier, if one searches for "*v*" using a search engine (e.g. Google and Baidu), then all the search results are $s(v) = F(v)$. *F(v)* can be considered as a similarity class of *v,* denoted as [*v*], that is, [*v*] = {*y* | *y* is a search result from searching for "*v*" online}. In other words, if $y_1, y_2 \in [v]$ then y_1 and y_2 are similar, denoted as $y_1 \approx y_2$. y_1 and y_2 are the search results from searching for "*v*" online. This example implies that two entities are similar *iff* they have something in common (here, they have the same *v*).

More generally, given an attribute value sequence $v_1, v_2, \cdots, v_n, \cdots$, then searching for each of them online, then $s(v_1) = [v_1] = F(v_1)$, $s(v_2) = [v_2] = F(v_2), \ldots, s(v_n) = [v_n] = F(v_n), \ldots$. If for any given $i \in \{1, 2, 3, \ldots, n, \cdots\}$ such that $v_i \approx v_{i+1}$, then $F(v_i) \cong F(v_{i+1})$. $y_1, y_2 \in F(v_i)$ are said to be infinite similar when the given *i* is sufficiently greater in *N*.

Infinite similarity originates from the experience of searching online using Google and our early work on similarity [32], when one searches for things on the web for many times and many years, s/he find that there is a kind of similarity among what searched appearing. This kind of similarity is infinitely similar.

Remark: This infinite similarity also reflects bounded rationality: individuals make decisions, their rationality is bounded by the available information, the tractability of the decision problem, the cognitive limitations of their minds, the time, environment, and technical conditions available to make the decision [4,33].

5.2 Weak and Strong Similarity

Given that sets $A_1, A_2, \cdots, A_n$ are an attribute sequence, where n is an integer. R is a relational database schema on these sets, denoted as $R(A_1, A_2 \cdots, A_n)$. Any instance of $R(A_1, A_2 \cdots, A_n)$ is a relation [19], that is, assume that $r \in R$ is a relation, there are n attributes $a_1, a_2 \cdots, a_n$ associated with r, that is, $r(a_1, a_2 \cdots, a_n)$ is an instance of $R(A_1, A_2 \cdots, A_n)$. A relation represents a fact, a piece of information, or a piece of knowledge. When $n = 0$, r is a fact. When $n = 1$, $r(a_1)$ represents a fact with an attribute a_1. For example, r(a) can represent "Paul is tall". r denotes " x is tall", a = Paul.

In a relational database [19,36],assume that the number of attributes $A_1, A_2 \cdots, A_n$ in a relation R is finite. For example, there are more than 300 attributes used for matching the love relationship between individuals on a matching website. In some cases, n > 1,000, in order to represent a piece of knowledge or information.

Now assume $R_1(A_{11}, A_{12} \cdots, A_{1N})$ and $R_2(A_{21}, A_{22} \cdots, A_{2N})$ are two relational schemas, where N is an integer.

Definition 3.

R_1 and R_2 are said to be similar with respect to relational schema, denoted as $R_1 \approx R_2$, iff there exists an integer $0 < K_1 \leq N$ so that for any $i \leq K_1$, $A_{1i} = A_{2i}$. In this case, we call R_1 and R_2 are similar with respect to a relational schema.

This concept of similarity is at a relational database schema level. It is useful for designing a relational database [19].

Definition 4.

In a relational database, assume that $r(a_1, a_2 \cdots, a_N)$ is a relation, an instance of $R(A_1, A_2 \cdots, A_N)$. $r(v_{i1}, v_{i2} \cdots, v_{iN})$ and $r(v_{j1}, v_{j2} \cdots, v_{jN})$ are row i and row j of r. then $r(v_{i1}, v_{i2} \cdots, v_{iN})$ and $r(v_{i1}, v_{i2} \cdots, v_{iN})$ are said to be K-similar, denoted as $r(v_{i1}, v_{i2} \cdots, v_{iN} \overset{K}{\approx} r(v_{j1}, v_{j2} \cdots, v_{jN})$ *iff* for a given K, $0 < K \leq N$, such that $v_{ik} = v_{jk}$, where $k = 1,2, \cdots, K$.

K-similarity is called as a weak similarity. It is weak because it is a kind of similarity at the record (row) level. This similarity is related to the data redundancy at the attribute level, and the record level, and therefore it is of practical significance in database management systems [19].

Example 2.

r(Sex, Program, ID, Name, Age) is a relational database, illustrated in the Table 1.

Table 1. A student relational database.

Sex	Program	ID	Name	Age
M	IT	160001	John	18
M	IT	160002	Peter	19
M	IT	160003	Lee	20
F	IT	160004	Liz	19
F	IT	160005	Lana	18
F	IT	160006	Bessie	19
F	IT	160007	Grace	20

Program oriented relation P = {<john, john>, <Peter, Peter>, <Lee, Lee> <John, Peter>, <Peter, John>, <John, Lee>, <Lee, John>, <Peter, Lee>, <Lee, Peter>; <Lana, Lana>, <Bessie, Bessie>, <Grace, Grace>, <Lana, Bessie>, <Bessie, Lana>, <Lana, Grace>, <Grace, Lana>, <Bessie, Grace>, <Grace, Bessie>}. Then P is a 2-similar relation. P partitions the above table into two similarity classes, [John] = {all the male IT students} and [Lana] = {all the female IT students}. This example implies that two male undergraduate students studying IT program at a university are similar.

Theorem 5.

K-similarity relation is a similarity relation.

Prove. To prove this theorem, it only needs to prove K-similarity relation is reflexive, symmetric, and transitive, based on Definition 2.

Assume that $r(v_{i1}, v_{i2} \cdots, v_{iN})$, $r(v_{j1}, v_{j2} \cdots, v_{jN})$ and $r(v_{h1}, v_{h2} \cdots, v_{hN})$ are row i, row j and row h of r, where $i, j, j\ h \in \{1, 2, ..., N\}$ .K is an integer, $0 < K \leq N$.

1) For $k \in \{1, 2, ..., K\}$, $v_{ik} = v_{ik}$, then $r(v_{i1}, v_{i2} \cdots, v_{iN}) \overset{K}{\approx} r(v_{i1}, v_{i2} \cdots, v_{iN})$, This means that K-similarity relation is reflexive.

2) If $r(v_{i1}, v_{i2} \cdots, v_{iN}) \overset{K}{\approx} r(v_{j1}, v_{j2}$, then for any k, $1 \leq k \leq K$, $v_{ik} = v_{jk}$ or $v_{jk} = v_{ik}$. Then $r(v_{j1}, v_{j2} \cdots, v_{jN}) \overset{K}{\approx} r(v_{i1}, v_{i2} \cdots, v_{iN})$ holds. That is, K-similarity relation is symmetric.

3) If $r(v_{i1}, v_{i2} \cdots, v_{iN}) \overset{K}{\approx} r(v_{j1}, v_{j2} \cdots, v_{jN})$ and $r(v_{j1}, v_{j2} \cdots, v_{jN}) \overset{K}{\approx} r(v_{h1}, v_{h2} \cdots, v_{hN})$, then for any $1 \leq k \leq K$, $v_{ik} = v_{jk}$, and $v_{jk} = v_{hk}$, that is, $v_{ik} = v_{hk}$, then $r(v_{i1}, v_{i2} \cdots, v_{iN}) \overset{K}{\approx} r(v_{h1}, v_{h2} \cdots, v_{hN})$ holds. Therefore, K-similarity relation is transitive.

This theorem demonstrates that weak similarity is a similarity relation.

When $K = N$, then a K-similarity relation is said to be strongly similar. It is easy to prove.

Theorem 6.

If $r(v_{i1}, v_{i2} \cdots, v_{iN})$ and $r(v_{j1}, v_{j2} \cdots, v_{jN})$ are strongly similar, then $r(v_{i1}, v_{i2} \cdots, v_{iN})$ and $r(v_{j1}, v_{j2} \cdots, v_{jN})$ are K-similar,

where $0 < K \leq N$.

Remark: Weak and strong similarities can facilitate saving, updating, and deleting data in a relational database. For example, if $r(v_{i1}, v_{i2}\cdots, v_{iN})$ and $r(v_{j1}, v_{j2}\cdots, v_{jN})$ are strongly similar, then one of the two rows (records) is redundant, and should be deleted. Furthermore, weak similarity can be considered a kind of partial similarity with respect to a row, whereas strong similarity is a kind of the whole similarity with respect to a row.

5.3 Database with Infinite Attributes and Similarity-based Reasoning for Big Data

This subsection applies the concept of limit to explore searching for big data based on the limit of number sequence in calculus [34]. It also delves into similarity-based reasoning for big data. More specifically, big data can be classified into two categories: One is database-based data, and another is NoSQL data [19]. Then this subsection discusses similarity-based reasoning for database-based big data and NoSQL big data. They are the basis for non-computation-based similarity, similarity-based infinite reasoning.

NoSQL databases such as Google's BigTable and Apache's Cassandra use the key-value data model or attribute-value model [19]. The attribute-value model consists of two data elements: an attribute and a value, in which every attribute has a corresponding value or a set of values. This can be considered as a simplified table different from the traditional tables that underpin the relational database. The simplified table in the NoSQL database consists of only three columns: AttributeID, Attribute, and Attribute value. For short, it is (AID, A, V), where AID is the ID of attribute $a \in A$, A is an attribute set, V is an attribute value set. The relationship between an attribute $a \in A$ and attribute value is 1:M, that is, an attribute $a \in A$ corresponds to a number of attribute values $v \in V$. Generally, N is the set of natural numbers. For any $i \in N, N = \{1, 2, 3, ..., n, ...\}$, $a_i \in A$ is an attribute, its attribute value is $ai(vj)$. When the cardinality of A equals to that of N, that is, $|A| = \aleph_0$, then (AID, A, V) is a relational database with infinite attributes.

A relational database with infinite attributes can be also defined as follows: let R be a relation. Its sequence of attributes is $A_1, A_2\cdots, A_n\cdots$, where n is an integer, $n \in N$. When n trends to infinity, $R(A_1, A_2\cdots, A_n\cdots,)$ is a relational database schema with infinite attributes. A relational database $r \in R(A_1, A_2 \cdots, A_n \cdots,)$, $r(a_1, a_2, \cdots, a_n, \cdots)$, is a relational database with infinite attributes *iff* it has a countable infinite attribute sequence $a_i \in A$.

Definition 5.

For any given integer K, if R_1 and R_2 are always similar with respect to relational schemas based on Definition 1, then we call R_1 and R_2 are infinitely similar with respect to a relational schema.

Definition 6.

Assume that $r(v_{i1}, v_{i2}\cdots, v_{in},\cdots,)$ and $r(v_{j1}, v_{j2}\cdots, v_{jn},\cdots,)$ are row i and row j of a relational database with infinite attributes, $r(a_1, a_2\cdots, a_n,\cdots)$. Then $r(v_{i1}, v_{i2}\cdots, v_{in},\cdots)$ and $r(v_{j1}, v_{j2}\cdots, v_{jn},\cdots)$ are said to be infinitely similar *iff* for any significantly big integer $K \in N$, $a_{ik} = a_{jk}$, where $k = 1, 2,\cdots, K$.

Theorem 7.

If $r(v_{i1}, v_{i2}\cdots, v_{in},\cdots)$ and $r(v_{j1}, v_{j2}\cdots, v_{jn},\cdots)$ are infinitely similar. Then they are K_1-similar for any $K_1 \leq K$.

From a practical viewpoint, only a few dozens of attributes or a few hundreds of attributes are not enough for characterizing an entity in the age of big data. This is the reason why we introduce a relational database with infinite attributes. The finite similarity in a relational database with infinite attributes paves the way from finite similarity to infinite similarity. This is useful for searching for big data and similarity-based search for a large database with infinite attributes. This is also useful for the development of human recognition because the practice in the finite world can be used to understand the infinite similarity in the infinite world.

6. A Logical Approach for Making Social Networking Services Big

Online social networking (OSN) services generate a big volume of big data. For example, YouTube generates 263 PB of big data yearly [17]. This section presents a logical approach to making social networking services (OSN) big as a part of applying mathematics to big data.

An OSN like Meta (Facebook) launched an application "people you may know" to directly (online) acquire email addresses based on your registered email address. The principle of this acquisition is illustrated in Figure 1.

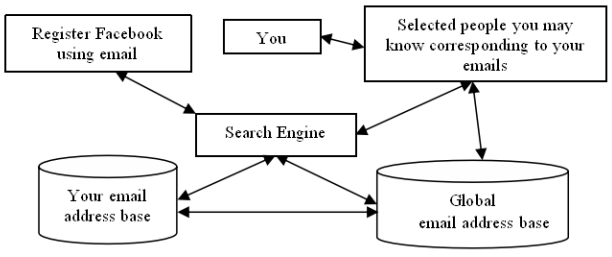

Figure 1. An intelligent technique of directly (online) acquiring email addresses based on a registered email address

As soon as one registered as a user of an OSN such as Meta (Facebook), WeChat, and LinkedIn, all of the email addresses in her or his email address base have been au-

tomatically exposed to the OSN. The OSN can automatically and regularly visit the registered email box, scan all the emails, extract information of email addresses that one has used, received, and sent, and then collect all of the email addresses away and store them in the Global email address base, as shown in Figure 1. Then the OSN can use any of the email addresses to contact the "friends" that s/he has used email to communicate with, the names have been in the email address base.

Friending is a marketing strategy of the OSN like Facebook and WeChat. It does not care if it is private to you. The friending mechanism of Facebook automatically invites your "friends" to join Facebook using "Selected people you may know". All these illustrated in Figure 1 are automatically realized using intelligent agents [2]. This is the reason why the leader of OSN or OSN services advises that you need not care about your privacy. If everyone opposes the invasion of his or her email address base for other purposes at the early time of Facebook, then Facebook would be disastrous. However, the social norm is just something that has evolved over time. One enjoys the services provided by OSN like Facebook and WeChat as well as TikTok, s/he has to sacrifice some privacy.

Assume that your email address base is E, F is your correspondence name base with respect to email communications, that is, for any name $f \in F$, there at least exists an email address $e \in E$, e is the email address of f. This means that any person that contacted you using email is your friend from a viewpoint of an OSN like Facebook, his or her email address is in E, and his or her name is in F. Now we have a virtual friendship, or email-based friendship as a binary relation, denoted as eF. eF is similar, because 1) You can email yourself. Then, eF is reflective. 2) If you can email anyone in E, then he or she can email you, then eF is symmetric. 3) If you can email your friend x, and your friend x can email his or her friend y, then you can email friend y, then eF is transitive.

The symmetry of eF makes everyone share the information in a symmetrical way. The transitivity of eF can make an OSN like Facebook market its services and acquire new customers, new Facebook friends. This is why an OSN can become globally popular within a short time.

A marketing strategy aims to acquire new customers, select customers, extend customers and retain customers profitably [4]. Now a logical approach to the marketing strategy of the OSN is presented below, based on Figure 1. Let : $P(f)$: $f \in F$ be a person. $N(e)$: e is an email address.

1) $P(f_0)$ (f_0 has registered as a member of OSN like Facebook)

2) $P(f_0) \rightarrow N(e_0)$ (f_0 submitted the email address e_0 to OSN)

3) $N(e_0)$ (The email address has been saved to OSN's global email address base) (1, 2) modus ponens

4) $\forall e(N(e_0) \rightarrow N(e))$ (Your contacted email address)

5) $N(e_0) \rightarrow N(e)$ (Remove the qualifier)

6) $N(e)$ (3, 5) (modus ponens)

7) $\forall e \exists f(N(e) \rightarrow P(f))$ (For any given email address, it corresponds to a person f)

8) $N(e) \rightarrow P(f)$ (Remove the qualifiers)

9) $P(f)$ (6, 8) (modus ponens)

10) $\exists f P(f)$ (Add the qualifier)

Then the OSN saves the information of $P(f)$ and tells you, "You may know $P(f)$". This logical approach implies that if f_0 is an OSN user, then the OSN can contact and attract all persons $P(f)$, $f \in F$, to become the *OSN* users. For example, if an average individual has 100 correspondence names. Then the OSN uses "You may know $P(f)$" to contact and attract all the corresponded persons five times one after another, exponentially, and then can attract $(100)^5 = (10)^{10}$ persons to become its users. Therefore, this automatic marketing approach brings about a bursting (exponential) increase of the OSN users just as Facebook has done in the past decade.

It should be noted that ResearchGate (https://www.researchgate.net/) has also used the technology based on the above-mentioned principle and logical approach. However, WeChat (www.wechat.com) have not mastered such a technology to attract its registered users to self-willingly expose their own privacy after submitting their own email address to WeChat. They still use traditional viral marketing for promoting their business. Viral marketing is based on a fact that a WeChat user invites his or her friends to use WeChat so that there are over 1 billion monthly active users in the WeChat world.

7. Related Work and Discussion

A number of scholarly research publications on big data have been mentioned in the previous sections. This section will focus on related work and discussion on a mathematical theory of big data.

Shannon's landmark article titled "A mathematical theory of communication" [35], provides the basis for information theory and has facilitated the lasting development of information science and technology since then. However, no articles titled a mathematical theory of big data have appeared so far. This inspires us to develop this article, which is an endeavor in this direction. This is also an extension and generalization of our early work from a mathematical foundation to a mathematical theory [9].

Google searches for "mathematics of big data" found about 32,100 results (27 November 2016, when this section was first written) and about 95,500 results (on April

20, 2022, when this section is updated). These results include courses offered by universities, workshops, and presentations on the mathematics of big data or data science. This means that mathematicians have paid attention to the dramatic development of big data and attempted to provide a mathematical approach to big data. For example, Laval has been developing a course on the mathematics of big data since 2015 [6,36]. The course provides students with mathematical techniques used to acquire, analyze, and visualize big data (e.g., using MATLAB) [37]. The workshop on mathematics in data science was held in 2015 in the USA [38]. Its objective is to explore the role of the mathematical sciences in big data as a discipline. Peter delivered a presentation on mathematics in data science at ICERM [7,41].

A Google scholar (www.scholar.google.com.au) searched for "mathematics of big data" in November 2016, when this section was first written, there were no searched article titles or book titles including "mathematics of big data". A Google scholar search for "mathematics of big data" was conducted on April 25, 2018 to update this research, and found that there were only 22 search results. A Google scholar search for "mathematics of big data" found 82 results on April 20, 2022. Four search results out of 22 are particularly worth discussing here. They are 1) Introduction to the Mathematics of Big Data [39,42]. 2) A Mathematical Foundation of Big Data [9]. 3) A Book on Applied Mathematics [40,43]. 4) The recently published book on mathematics of big data [44].

The first is a course description for the course with the same name. This course has been offered since 2015 [37,39,42]. It gives a short overview of big data and discusses the issues associated with big data with some answers.

The second [9] examines big as an operation, the cardinality of big data, and explores a mathematical approach to searching big data. However, the work of Sun and Wang [9] lacks logical approach and other mathematical approaches that are necessary for developing a mathematical theory of big data. This article updates and generalizes some of its results, and further explores infinite similarity and logical approach to online social networking platforms.

The third states that the mathematics of big data can provide theories, methods, and algorithms for processing, transmitting, receiving, understanding, and visualizing datasets [40,43].

The last is a book focusing on applications and practices of spreadsheets, databases, matrices, linear algebra, and graphs and for processing big data [44].

Big data is a market-inspired brand and research field. It seems to lack rigorous research from a perspective of mathematics. This is similar to social computing which "benefits from mathematical foundations, but research has barely scratched the surface" [45]. The above discussion implies that there is still a long way to go to develop the mathematics of big data as a discipline. This article provides an attempt to explore a mathematical theory for big data based on the work of Sun and Wang [9] and motivated by C. E. Shannon [35]. More theoretical work will be undertaken to develop a mathematical theory of big data and big data analytics.

Fuzzy sets and fuzzy logic [29,32] have been used to explore the relativity of big data and showed that one should have inclusiveness in exploring big data so that everyone can get benefit from the research and development of big data with applications. Furthermore, two big characteristics of big data are big volume and big veracity [5]. The big volume of big data is fuzzy in essence. The big veracity is related to the ambiguity and incompleteness of big data [46]. Fuzzy logic and fuzzy sets have developed a significant number of methods and techniques to address ambiguity and incompleteness of data, and therefore they will play a significant role in overcoming ambiguity and incompleteness of big data [30,32,47].

8. Conclusions

The objective of this article is to apply mathematics to treat a few fundamental problems of big data and develop a mathematical theory of big data. To this end, it explores the volume of big data with the cardinality theory. It provides a mathematical foundation for searching for big data with the set theory. It reveals the relativity of big data with fuzzy logic and fuzzy sets theory [28]. It presents a similarity-based approach to big data by investigating finite and infinite similarity, the weak and strong similarity of big data, and similarity-based infinite reasoning. It also presents a logical approach to marketing strategy for online social networking platform services. The research contributes to the literature along three dimensions: 1) Cardinality of big data is the same as the cardinality of all the real numbers. 2) The relativity and infinity are two big characteristics of big data besides the ten big characteristics of big data [8]. The relativity of big data leads to the continuum from small data to big data, big data-driven small data analytics becomes statistically significant for further research and development of big data [14]. The infinity of big data leads to the exploration of infinite similarity of big data. 3) A logical foundation for revealing the secret behind the success of Meta (Facebook) and other social networking platforms or services will lead to logical methods for big data and big data analytics besides machine learning and deep learning [2,3]. The proposed

approach in this article might facilitate the research and development of big data, big data analytics, big data computing, data science, and data intelligence.

The mathematic theory for big data, analytics, and processing is a very important issue that is worthy of paying great attention to study. A mathematical theory of big data should also include addressing the following questions: What is a fuzzy-logic theory of big data? What is a similarity-based theory of big data? What is a calculus of big data? What is the cyclic model of big data reasoning? All these require further deep investigation in the near future. We will present the calculus of big data, the calculus of analytics, and big data reasoning as research results soon.

Optimization has drawn increasing attention in the field of big data in general and big data analytics in particular [22], because it is the foundation of big data predictive analytics in general and big data prescriptive analytics. In future work, we will examine the process of optimization for big data descriptive analytics taking into account the life cycle of business process-oriented big data analytics.

Conflict of Interest

There is no conflict of interest.

References

[1] Sun, Z., Wu, Z., 2021. A Strategic Perspective on Big Data Driven Socioeconomic Development. in The 5th International Conference on Big Data Research (ICBDR).September 25-27 (pp. 35-41). Tokyo, Japan: ACM.

[2] Russell, S., Norvig, P., 2020. Artificial Intelligence: A Modern Approach (4th Edition), Upper Saddle River: Prentice Hall.

[3] Hurley, R., 2019. Data Science: A Comprehensive Guide to Data Science, Data Analytics, Data Mining, Artificial Intelligence. Machine Learning, and Big Data, Middletown, DE: Hurley.

[4] Laudon, K.G., Laudon, K.C., 2020. Management Information Systems: Managing the Digital Firm (16th Edition), Harlow, England: Pearson.

[5] Sun, Z., 2022. A Service-Oriented Foundation for Big Data. Research Anthology on Big Data Analytics, Architectures, and Applications, Hershey, PA, IGI-Global. pp. 869-887.

[6] Laval, P.B., 2015. The Mathematics of Big Data.

[7] Peters, T.J., 2015. Mathematics in Data Science.

[8] Sun, Z., Strang, K., Li, R., 2018. Big data with ten big characteristics. Proceedings of 2018 The 2nd Intl Conf. on Big Data Research (ICBDR 2018). October 27-29 (pp. 56-61). Weihai, China: ACM.

[9] Sun, Z., Wang, P.P., 2017. A Mathematical Foundation of Big Data. Journal of New Mathematics and Natural Computation. 13(2), 8-24.

[10] Sun, Z., Xiao, J., 1994. Essentials of Discrete Mathematics, Problems and Solutions., Baoding: Hebei University Press.

[11] Johnsonbaugh, R., 2013. Discrete Mathematics (7th Edition), Pearson Education Limited.

[12] Enderton, H., 1977. Elements of Set Theory, Academic Press Inc.

[13] McAfee, A., Brynjolfsson, E., 2012. Big data: The management revolution. Harvard Business Review. 90(10), 61-68.

[14] Sun, Z., Huo, Y., 2021. The spectrum of big data analytics. Journal of Computer Information Systems. 61(2), 154-162.

[15] Sallam, R., Friedman, T., 2022. Top Trends in Data and Analytics.

[16] Minelli, M., Chambers, M., Dhiraj, A., 2013. Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses, Wiley & Sons (Chinese Edition 2014).

[17] National Research Council, 2013. Frontiers in Massive Data Analysis, Washington, DC: The National Research Press.

[18] Clissa, L., 2022. Survey of Big Data sizes in 2021. (Online). Available: https://arxiv.org/abs/2202.07659. (Accessed 11 March 2022).

[19] Chen, P.P., 1976. The Entity-Relationship Model-Toward a Unified View of Data. ACM Transactions on Database Systems. 1(1), 9-36.

[20] Coronel, C., Morris, S., Rob, P., 2020. Database Systems: Design, Implementation, and Management (14th edition), Boston: Course Technology, Cengage Learning.

[21] Courant, R., 1961. Differential and Integral Calculus Volume I, Glasgow: Blackie & Son, Ltd.

[22] Kelly, J.E., 2015. Computing, cognition and the future of knowing.

[23] Sun, Z., Pambel, F., Wu, Z., 2022. The Elements of Intelligent Business Analytics: Principles, Techniques, and Tools. Handbook of Research on Foundations and Applications of Intelligent Business Analytics, Z. Sun and Z. Wu, Eds. pp. 1-20.

[24] Halevy, A., Norvig, P., Pereira, F., 2009. The Unreasonable Effectiveness of Data. IEEE Intelligent Systems. pp. 8-12.

[25] Jech, T., 2003. Set Theory: The Third Millennium Edition, Revised and Expanded., Springer.

[26] Manyika, J., Chui, M., Bughin, J.E.A., 2011. Big data:

The next frontier for innovation, competition, and productivity. (Online). Available: http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation

[27] Sharda, R., Delen, D., Turban, E., et al., 2018. Business Intelligence, Analytics, and Data Science: A Managerial Perspective (4th Edition), Pearson.

[28] Lang, S., 2002. Algebra, Graduate Texts in Mathematics 211 (Revised third ed.), New York: Springer-Verlag.

[29] Zimmermann, H., 2001. Fuzzy set theory and its applications (4th edition), Boston: Kluwer Academic Publishers (Springer Seience+Business Media New York).

[30] Zadeh, L.A., 1979. Fuzzy sets and information granularity. Advances in Fuzzy Sets Theory and Applications, Horth-Holland, New York, Elsevier. pp. 3-18.

[31] IGI, 2015. Big Data: Concepts, Methodologies, Tools, and Applications.

[32] Zadeh, L.A., 1965. Fuzzy sets. Information and Control. 8(3), 338-353.

[33] Sun, Z., Sun, L., Strang, K., 2018. Big Data Analytics Services for Enhancing Business Intelligence. Journal of Computer Information Systems (JCIS). 58(2), 162-169.

[34] Finnie, G., Sun, Z., 2002. Similarity and metrics in case-based reasoning. International Journal Intelligent Systems. 17(3), 273-287.

[35] Gigerenzer, G., Selten, R., 2002. Bounded Rationality: The Adaptive Toolbox., MIT Press.

[36] Sun, Z., Pinjik, P., Pambel, F., 2021. Business case mining and E-R modeling optimization. Studies in Engineering and Technology. 8(1), 53-66.

[37] Larson, R., Edwards, B.H., 2010. Calculus (9th ed.), Brooks Cole Cengage Learning.

[38] Shannon, C.E., 1948. A mathematical theory of communication. The Bell System Technical Journal. 27, 379-423, 623-656.

[39] Laval, P.B., 2015. MATH 7900/4490 Math The Mathematics of Big Data (Syllabus). [Online]. Available: https://math.kennesaw.edu/~plaval/BigData/syllabus.pdf. (Accessed 4 Sept 2016).

[40] Laval, P.B., 2015. Introduction to the Mathematics of Big Data.

[41] ICERM, 2015. Mathematics in Data Science. (Online). Available: https://icerm.brown.edu/topical_workshops/tw15-6-mds/

[42] Laval, P.B., 2017. Introduction to the Mathematics of Big Data. (Online). Available: http://ksuweb.kennesaw.edu/~plaval/math4490/fall2017/mathsurvey_def.pdf. (Accessed 25 4 2018).

[43] Chui, C.K., Jiang, Q., 2013. Applied Mathematics: Data Compression, Spectral Methods, Fourier Analysis, Wavelets, and Applications, Springer.

[44] Kepner, J., Jananthan, H., 2018. Mathematics of Big Data: Spreadsheets, Databases, Matrices, and Graphs, MIT Press.

[45] Chen, Y., Ghosh, A., Kearns, M., 2016. Mathematical foundations for social computing. CACM. 59(10), 102-108.

[46] IBM, 2015. The Four V's of Big Data. (Online). Available: http://www.ibmbigdatahub.com/infographic/four-vs-bigdata

[47] Kantardzic, M., 2011. Data Mining: Concepts, Models, Methods, and Algorithms, Hoboken, NJ: Wiley & IEEE Press.